ScicomP13 2007
SP-XXL

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Monitoring Infrastructure for Superclusters: Experiences at MareNostrum

Garching, Munich

Ernest Artiaga
Performance Group
BSC-CNS, Operations

Outline

- BSC-CNS and MareNostrum Overview
- Monitoring Requirements
- Building Blocks for a Monitoring System
- GGcollector: Architecture and Implementation
- Lessons Learned and Future Work

The BSC-CNS Consortium

- The BSC-CNS Consortium includes:
  - Spanish Government (MEC)
  - Catalonian Government (DURSI - Generalitat)
  - Technical University of Catalunya (UPC)

- Started operations on 2005
- Located at UPC North Campus

Generalitat de Catalunya
Departament d'Universitats, Recerca i Societat de la Informació

MINISTERIO DE EDUCACION Y CIENCIA

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Top500

TOP500® CERTIFICATE

MareNostrum, an IBM eServer BladeCenter JS21 System at the Barcelona Supercomputing Center, Spain

is ranked

No. 1 in Europe

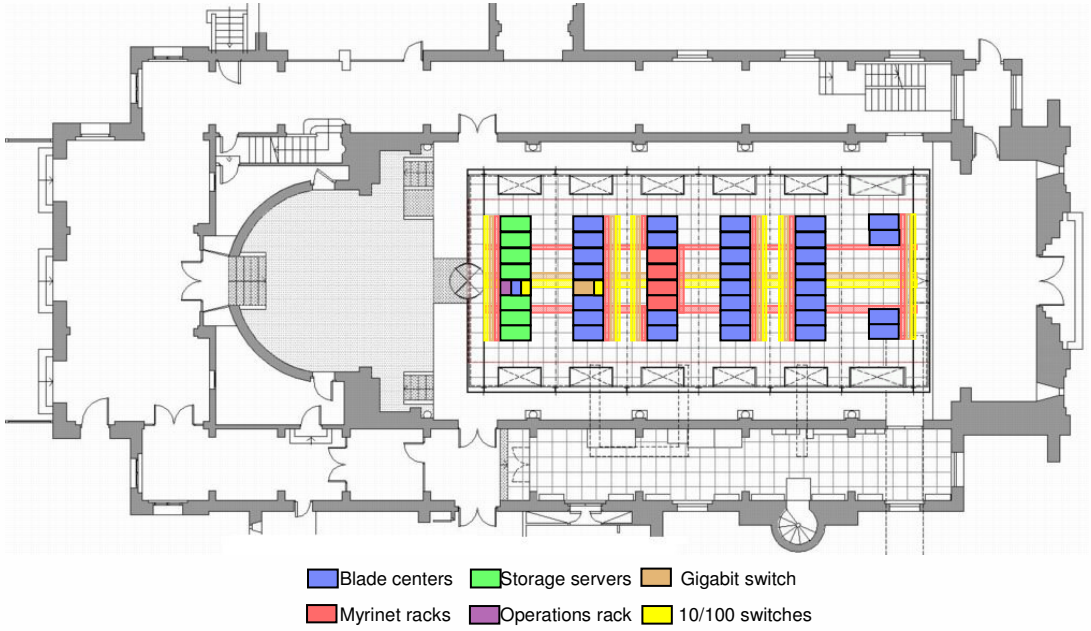
among the World's TOP500 Supercomputers
with 62.63 TFlop/s Linpack Performance
on the TOP500 List published at the SC06 Conference, November 14, 2006

Congratulations from the TOP500 Editors

Hans Meuer	Erich Strohmaier	Jack Dongarra	Horst Simon
University of Mannheim	NERSC/Berkeley Lab	University of Tennessee	NERSC/Berkeley Lab

MareNostrum Overview

- Computing Nodes
  - 2560 JS21, 2.3GHz
  - 4 cores per board
  - 8 Gbytes RAM
  - 36 Gbytes disk
- Network
  - Myrinet
    - 2 Spine 1280
    - 10 Clos256
    - 2560 Myrinet cards
  - Gigabit Ethernet
  - 10/100 Ethernet



94,20 peak TFlops
10240 processors
20 TBytes memory
280 + 90 TBytes disk

MareNostrum Overview

- Novell SuSE Linux (kernel 2.6) on each node
- OS image loaded via Gigabit network
- Batch System
  - Slurm + Moab (Central Manager node)
- File systems
  - NFS & GPFS provided by 41 storage servers
  - Local disk for scratch

Outline

- BSC-CNS and MareNostrum Overview
- Monitoring Requirements
- Building Blocks for a Monitoring System
- GGcollector: Architecture and Implementation
- Lessons Learned and Future Work

Supercluster Specifics

- Increased risk of error conditions
  - Applications span across hundreds (thousands) of nodes
  - Probability of hardware/software/network component failure
- Heterogeneous workload
  - In practice, general purpose means “no patterns”
  - Application/system behaviour changes with individual projects
  - Per-processor allocation vs. per-node allocation
- Goals
  - Detect, process and report component malfunctions
  - Analyze the performance of MareNostrum
  - Support for diagnostic and decision-making

Requirements

- Minimize impact on applications
  - Small nodes need light probes
- High scalability
  - Big clusters do not like broadcasts and synchronisms
- Heterogeneous data sources
  - Operating system and scripts on each node
  - Management modules (from “blade center” technology)
  - Storage and external server logs and tools
  - Batch system controller, job prologs and epilogs
  - Network equipment
- *Redundant information to find discrepancies*

Flexibility

- Different data views for humans and systems
  - Overall, synthetic high-level view for system administrators
  - Detailed status for diagnostics
  - Historical records for performance and management
  - Batch system interface for taking scheduler decisions
  - Low-level format for alert systems
    - automatic preventive/corrective actions
- Able to deal with new software
  - Take advantage of existing tools
- Location independent
  - Decouple data acquisition/processing/storage/view

Outline

- BSC-CNS and MareNostrum Overview
- Monitoring Requirements
- Building Blocks for a Monitoring System
- GGcollector: Architecture and Implementation
- Lessons Learned and Future Work

The GGCollector Framework

- Good monitoring tools already exist, but…
  - Do not fulfil all our requirements (mainly scalability/footprint)
  - Difficult interoperation with other tools
  - Though some components of the tools may be nice!
- Additionally, we would like a single tool for everything
  - At least a common interface
- Our solution: GGCollector
  - A framework to integrate other tools' components
  - Provides collection/storage/view independence
  - Compensates weak points from other tools

Building Blocks: Ganglia

- http://ganglia.sourceforge.net/
  - Widely used monitoring system
- Three components
  - Gmond: per-node information
  - Gmetad: central collector for all cluster data
  - Web interface: visualization tool
- Interesting features
  - Gmond's small footprint and extensible metrics
  - Easy to parse XML data exchange
  - Poor scalability on big flat networks
  - Visualization tightly coupled with data collection

Building Blocks: RRDTools

- http://oss.oetiker.ch/rrdtool
  - Common tool for handling historical data
  - Based on large circular buffers
- Features
  - Can handle several data sources
  - Provides simple statistical functions for collected data
  - Limited source aggregation capacity (scalability limits)
  - Requires predefined fixed-size time slices
- Lots of front-ends
  - Including web-based graphical interfaces like drraw
    - http://web.taranis.org/drraw/

Building Blocks: Other Sources

- Batch system
  - Job wrapper scripts on nodes
  - Controller daemon data on head server
  - Data available via API
- Air conditioning
  - Chiller machines status from ad-hoc scripts
- More to come...
  - Management module data (node temperature, status, …)
  - Network equipment (interface to mrtg)
    - http://oss.oetiker.ch/mrtg

Outline

- BSC-CNS and MareNostrum Overview
- Monitoring Requirements
- Building Blocks for a Monitoring System
- GGcollector: Architecture and Implementation
- Lessons Learned and Future Work

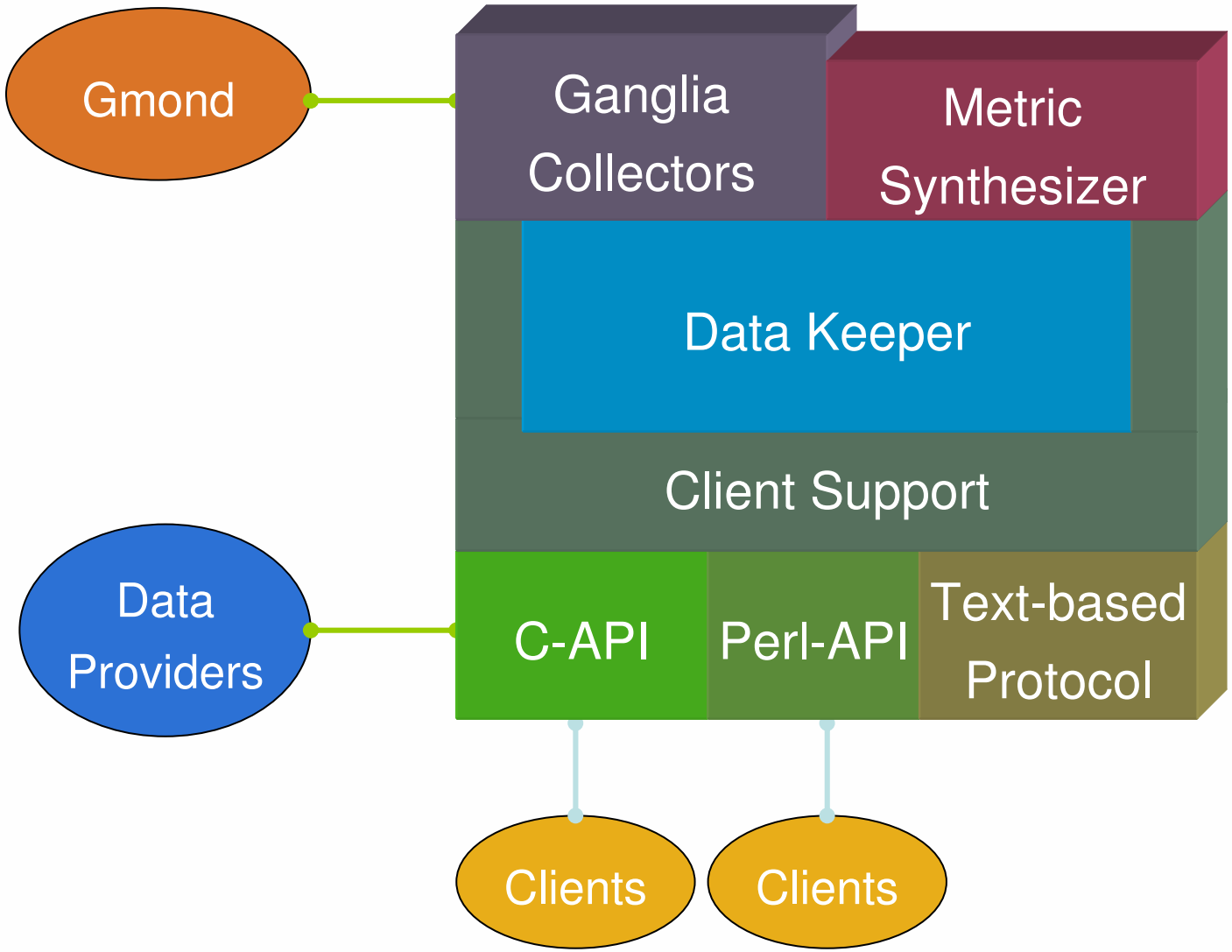
GGCollector

- Core of the MareNostrum monitoring infrastructure
  - Developed at BSC-CNS
  - Framework to integrate/complement other tools
  - Avoid unnecessary data movements
- Metric: basic piece of data
  - Applicable host (may differ from producer)
  - Time stamp
  - Type
  - Value
- Metric type attributes
  - Data type (integer/float/string)
  - Ttl (expiration time)
  - Threshold (value changes relevant for reporting)
  - Source

GGCollector

- Metric Synthesizer
  - Scalable aggregation facility for data from different nodes
  - Provides simple operations: count, average, max, min, sum.
- Subscription mode
  - Clients are notified only when relevant changes occur
  - Only requested data is sent
  - Full “query” mode is also available for one-time requests
- Other features
  - Low resource consumption
  - Basic CLI interface (extensible via API)
  - Distributed client/server model
  - Scalable by “chaining” (either as proxy or peer)

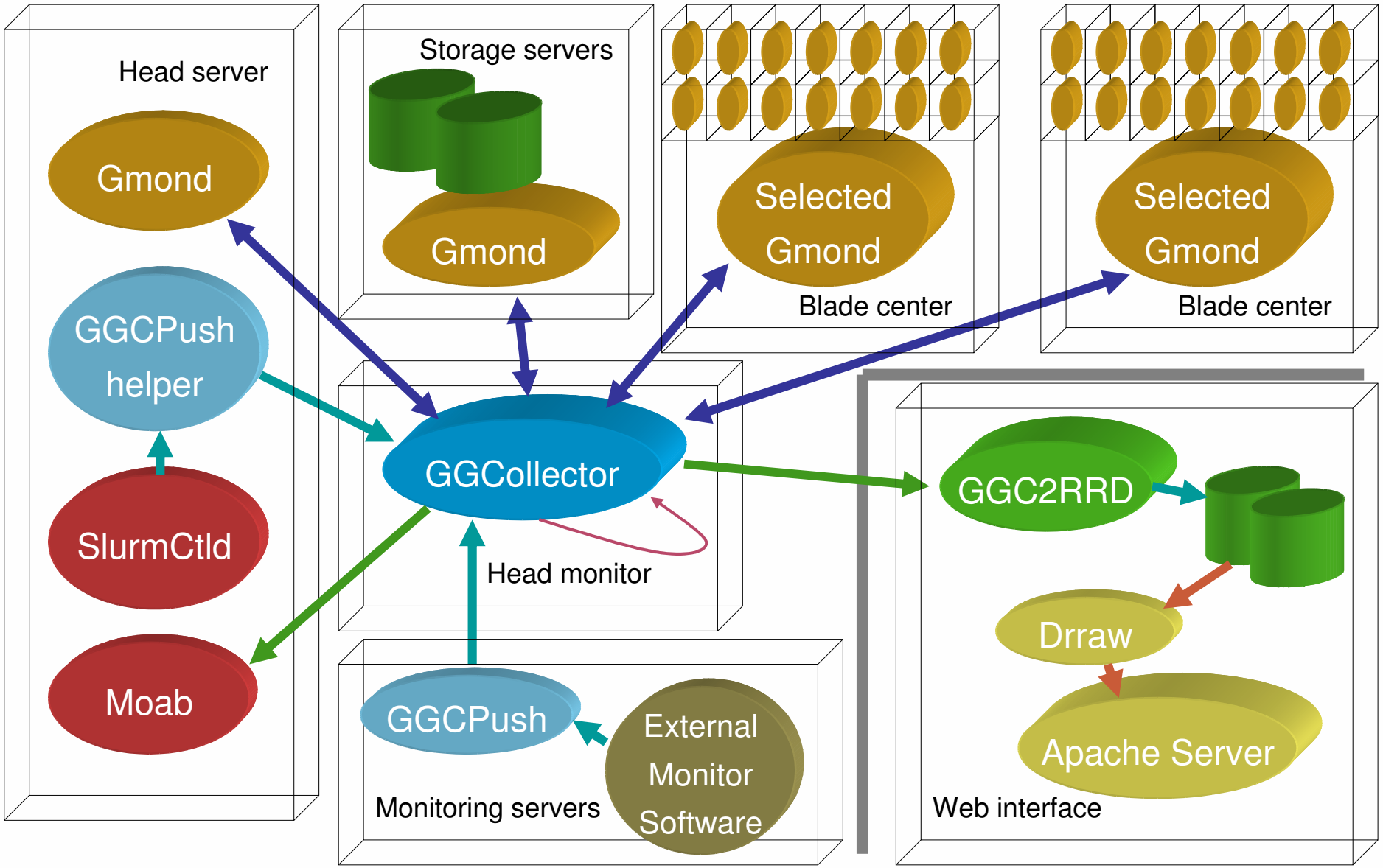
GGcollector: Architecture



GGCollector: Configuration

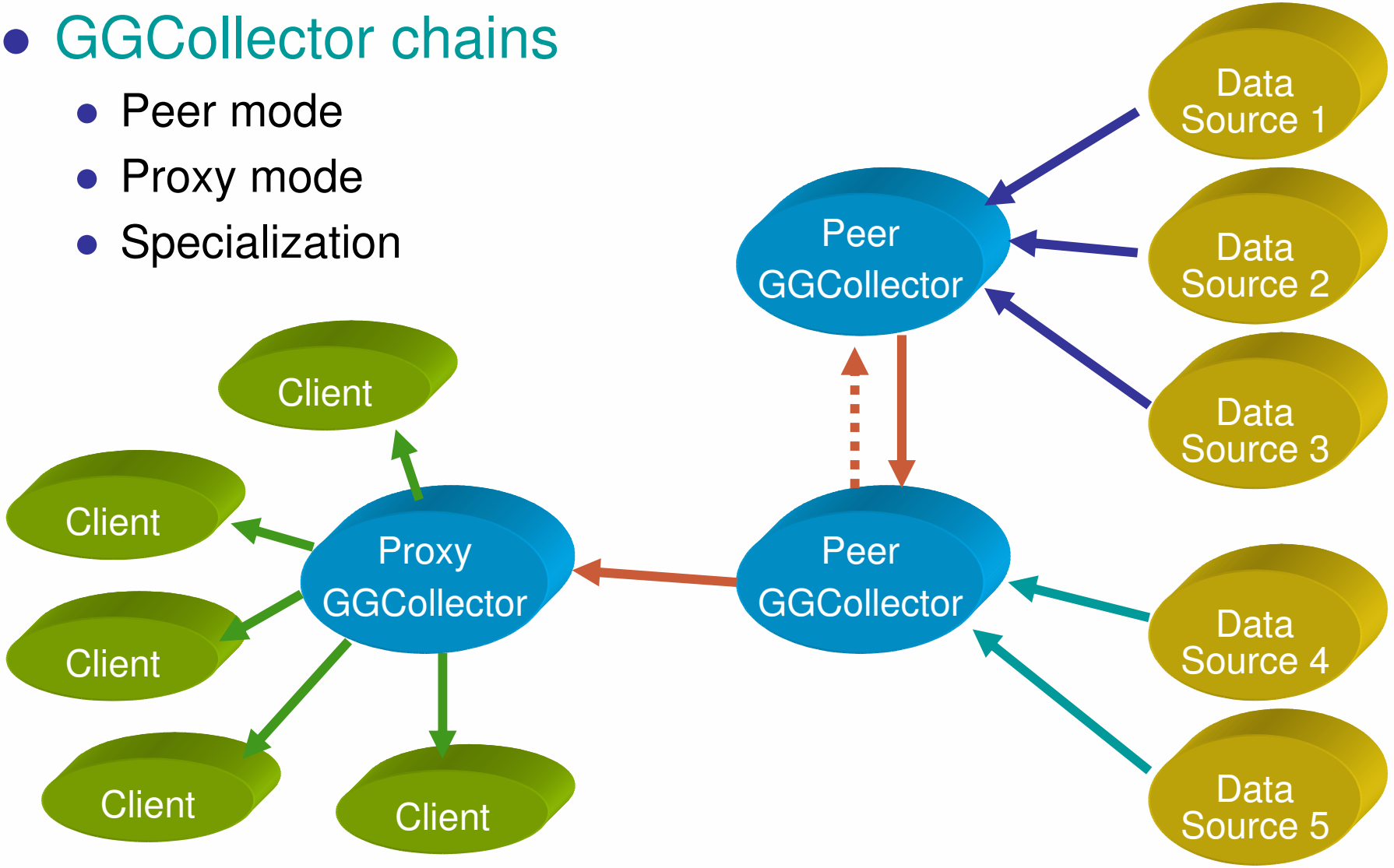
- Ganglia configuration
  - No Gmetad
  - Data shared only inside a Blade Center
- GGCollector configuration
  - Both metric types and applicable hosts are pre-defined
  - Tuneable number of threads
- Other sources
  - Use standard GGCPush client
  - Need data translation into GGCollector format
- Data acquisition
  - When necessary
  - Opposed to “As Fast As Possible”

GGCollector: Deployment



GGCollector: Scalability

- GGCollector chains
  - Peer mode
  - Proxy mode
  - Specialization



Examples

- http://www.bsc.es

CONTACT | QUICK LINKS | SITEMAP Search

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MareNostrum

A cutting-edge facility at the service of research, knowledge and development

ABOUT BSC | COMPUTER SCIENCES | EARTH SCIENCES | LIFE SCIENCES | MARENOSTRUM

Home > MareNostrum

- Access MN Form
- Mobility programs
- Services
- System Architecture

Related links

Distributed European Infrastructure for Supercomputing Applications

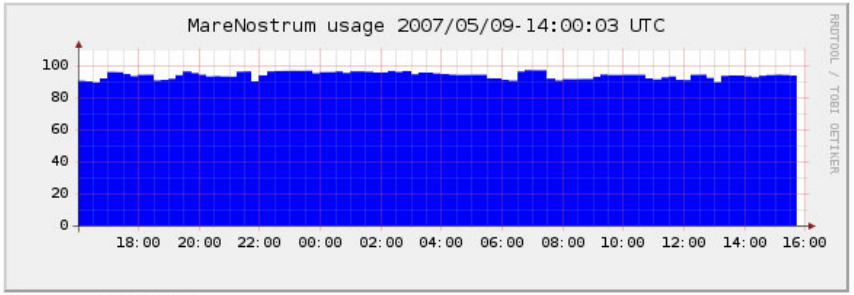
- Gallery

Job Offers

- Research Scholarships for the DEISA Project (Operations)
- System Administration

MareNostrum

BSC-CNS hosts MareNostrum, the most powerful supercomputer in Europe and the fifth in the world, according to the last Top500 list. In March 2004 the Spanish government and IBM signed an agreement to build one of the the fastest computer in Europe. In November 2006 its capacity has been increased due to the large demand of scientific projects. MareNostrum has increased the calculation capacity of the supercomputer MareNostrum, until reaching 94.21 Teraflops (94.21 trillions of operations per second), doubling its previous capacity (42.35 Teraflops). It had 4.812 processors and has now 10.240 processors with a final calculation capacity of 94.21 Teraflops.



MareNostrum current workload

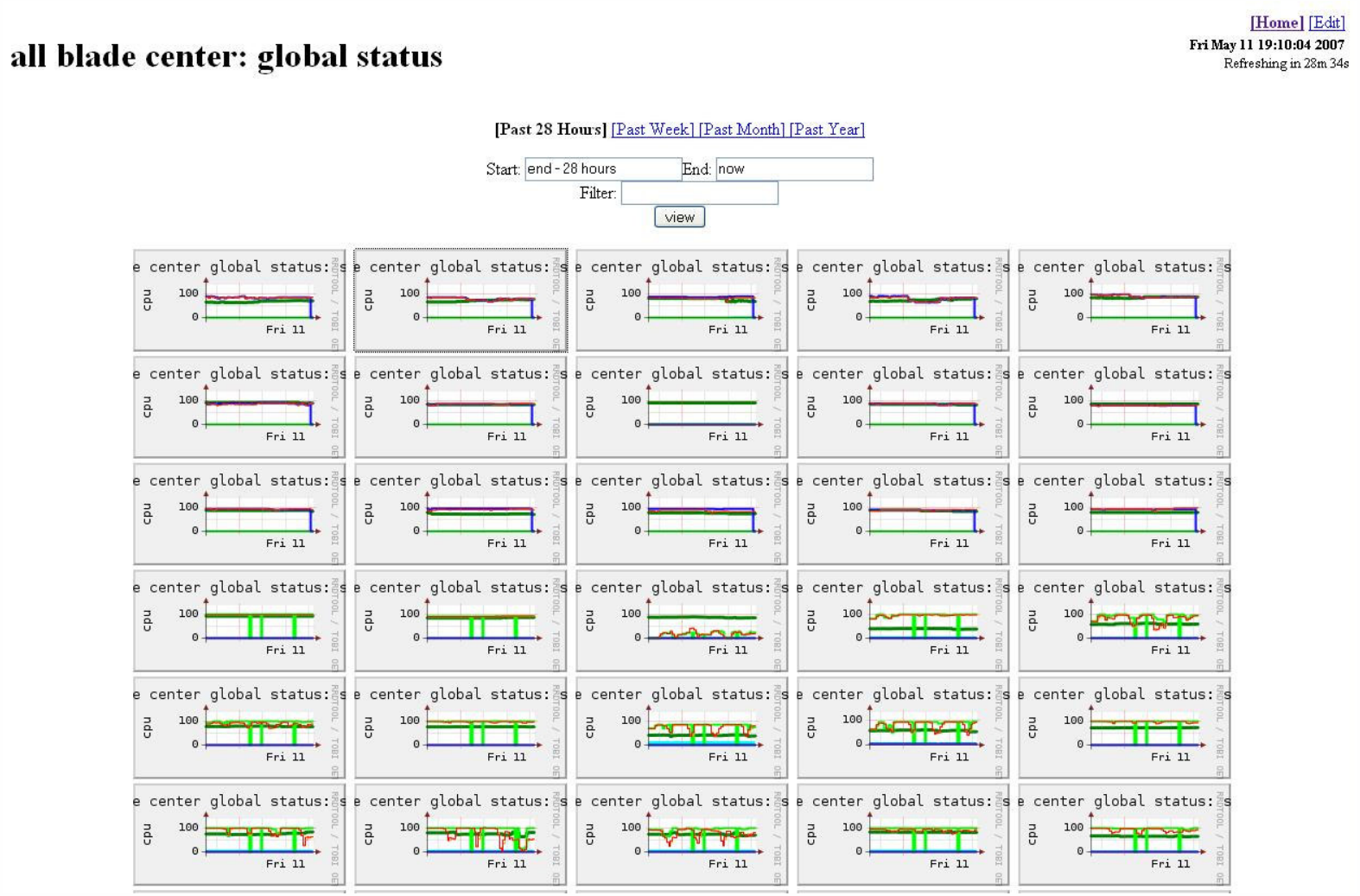
MareNostrum is managed by the Operations team that takes care of its availability, security and performance. An important task of this team is to support scientists in the usage of MareNostrum, as well as to help them in the improvement of their applications getting better research results.

These resources and expertise are not just available remotely. Spanish scientists as well as European ones can visit BSC-CNS through the available mobility programs in order to work together with our experts in supercomputing and learn more about how to improve their work and research results.

- **MareNostrum application form**: Please submit this form in order to request access for the different center resources such as

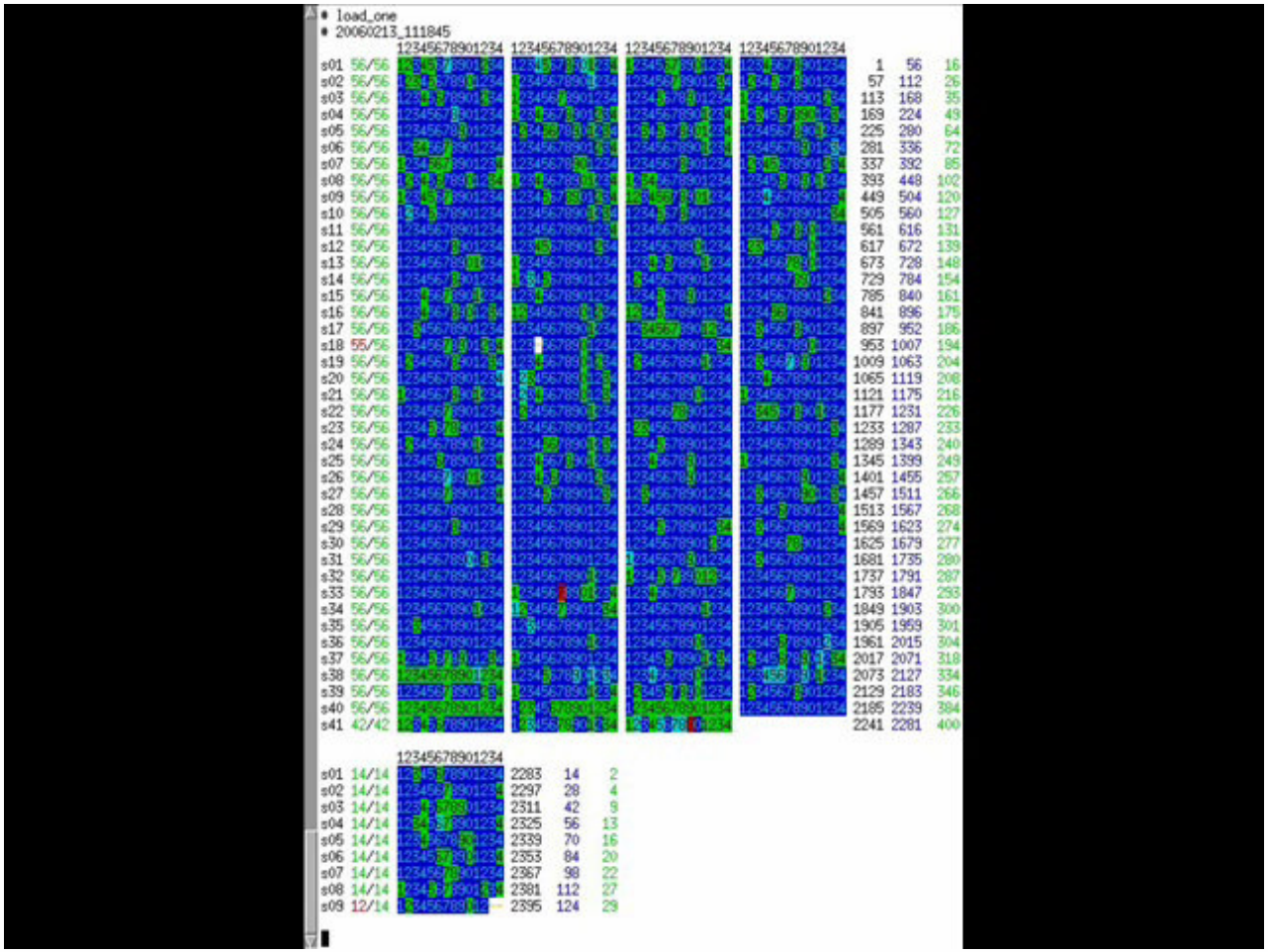
Examples

- Internal usage screen with ddraw



Examples

- Ggstatus: command line tool (text based).



Outline

- BSC-CNS and MareNostrum Overview
- Monitoring Requirements
- Building Blocks for a Monitoring System
- GGcollector: Architecture and Implementation
- Lessons Learned and Future Work

Lessons Learned

- Minimize impact on running applications
  - Tools with small footprint (memory and cpu) on nodes
  - Collect information only when necessary
    - Delays are acceptable!!!
- Hardware resources for monitoring are limited
  - Move the data out and process elsewhere
  - Decouple data acquisition/processing/storage/view
- Scalability: think thousands, not hundreds!
  - Avoid synchronisms
  - Avoid broadcasts
- Simple, generic, extensible framework
  - Easy to integrate, easy to extend

Future Work

- GGCollector is currently in production at BSC-CNS
  - Holds together the MareNostrum monitoring infrastructure
- Still room for improvement
  - Handling of thresholds and timeouts (per client?)
  - Improvement of chained configuration
- Integration with other tools
  - Take advantage of existing software
  - Nagios alert system (http://nagios.org)
  - Mrtg (http://oss.oetiker.ch/mrtg)
  - Database backend for historical data
- Application performance analysis
  - Crossing collected data with batch system job information

Spanish Supercomputing Network

GGCollector deployment

MareNostrum

Process: 10240 PowerPC 970 2.3 GHz
Memory: 20 TBytes
Disk: 280 + 90 TBytes
Network: Myrinet, Gigabit, 10/100
System: Linux

CeSViMa

Process: 2408 PowerPC 970 2.2 GHz
Memory: 4.7 TBytes
Disk: 63 + 47 TBytes
Network: Myrinet, Gigabit, 10/100
System: Linux

IAC, UMA, UNICAN, UNIZAR, UV

Process: 512 PowerPC 970 2.2 GHz
Memory: 1 TByte
Disk: 14 + 10 TBytes
Network: Myrinet, Gigabit, 10/100
System: Linux

Thank you !

www.bsc.es